Cognitive bias in animal behavior science: A philosophical perspective

Behzad Nematipour*, Marko Bračić, Ulrich Krohs

Affiliations

BN: Center for Philosophy of Science, University of Münster, Domplatz 23, 48143 Münster, Germany
MB: Department of Behavioural Biology, University of Münster, Badestr. 13, 48149 Münster, Germany
UK: Department of Philosophy, University of Münster, Domplatz 23, 48143 Münster, Germany

Correspondence

Behzad Nematipour, behzad.nematipour@uni-muenster.de

Abstract

Emotional states of animals influence their cognitive abilities as well as their behavior and welfare. Assessing emotional states is, therefore, indispensable for animal welfare science as well as for many fields of neuroscience, animal behavior science etc. This can be done in different ways. This paper focuses on the so-called *cognitive judgment bias* test, which has gained special attention in the last two decades and became a highly important way of measuring emotional states in non-human animals. However, less attention has been given to the epistemology of the cognitive judgment bias test and to disentangling the relevance of different steps in the underlying cognitive mechanisms. This paper sheds some light on both, the epistemology of the methods, and the architecture of the underlying cognitive abilities of the tested animals. Based on this reconstruction, we propose a scheme for classifying and assessing different cognitive abilities involved in cognitive judgment.

Keywords: Ambiguous stimuli; Cognitive bias; Emotions; Representation; Decision making

1 Introduction

Assessing animals' emotional states has explanatory, predictive and illustrative value for animal welfare science and in areas like neuroscience and psychopharmacology (Mendl et al. 2009) as well as for attributing rights to 'sentient species'. However, this assessment is particularly difficult in non-human animals because of the lack of verbal communication – or so called "linguistic reports". That is why scientists in these fields are looking for various

indicators of emotional states such as behavioral and physiological changes that accompany such states in order to assess in which emotional state an animal is, or whether or not animals of the considered species have them at all (Kremer et al., 2020). For example, the state of fear may be accompanied, in various animals, by behavior like freezing, fleeing and even attacking and by physiological changes such as increased heart rate, blood pressure and enhanced levels of circulating glucocorticoids (Mendl et al. 2009). But because of certain epistemic problems with these types of indicators (which are discussed in detail in the next section), scientists are inclined to consider other and potentially more reliable indicators (Kremer et al., 2020).

An increasingly used indicator of emotional states in non-human animals is "cognitive bias" (Paul et al. 2005). This indicator has its background in psychological experiments on humans: emotional states in humans affect their information processing and their judgements. A paradigmatic example of such influences is that people in negative emotional states, like anxiety, depression, or fear, make overwhelmingly often negative judgments about events or interpret ambiguous situations negatively. Although there are clear behavioral indications of anxiety, depression, fear etc. in non-human animals, especially in mammals, for example contact avoidance or self-isolation, it was until recently less clear whether or not these emotional states influence, like in humans, 'decision making' and 'judgments' about, or evaluation of, situations and events.

Potential utility of testing "cognitive bias" in welfare research was demonstrated in the seminal study of Harding et al., who demonstrated that rats housed in 'unpredictable'/stressful conditions (which causes depression-like symptoms) were inclined to respond more negatively to ambiguous situations than rats that were housed in 'predictable'/familiar conditions. Their judgment was biased (Harding et al. 2004).[1] The authors suggested to use behavioral responses in ambiguous situations as an indicator of emotional state, which initiated numerous studies that demonstrated that cognitive judgment bias can be found in a wide range of taxa, from pigs to bumblebees (e.g., Harding et al. 2004; Paul et al. 2005; Mendl et al. 2009; Lagisz et al., 2020; Neville et al., 2020). This led the scientists to believe that they can use this so called "cognitive judgment bias test" to measure emotional states of non-human animals.

In this paper we pursue two main goals. First, we want to examine the epistemic role of emotional indicators and their limits. We start by pointing at epistemic problems with the

[1] What exactly the housing conditions were and what it meant to „respond more negatively to ambiguous situations" will be clarified and discussed later in the paper.

more traditional indicators of emotional states (behavioral and physiological changes) and then see what the advantages and limits of 'cognitive bias' as a rather new indicator of emotional states in animals are. We aim at assessing the epistemic value of the cognitive judgment bias test and demonstrate its empirical motivation.

Second, we scrutinize cognitive bias as such. What kind of cognitive abilities/processes are in play? We are not engaging in a conceptual analysis of the notion of cognitive bias, but rather look at cognitive abilities/processes underlying the judgment bias that is used as an indicator of emotional states and aim at determining what kind of abilities/processes these are. While animal-welfare-science might not need to determine exactly what kind of ability is used as an indicator as long as there are proper ways of tracking or individuating them. However, from other perspectives this question is worth pursuing, because 1) for cognitive science and philosophy of mind, the (exact) kind of cognitive abilities of non-human animals is germane to understanding higher cognitive abilities and language acquisition in humans, and in an evolutionary perspective as well for non-human animals. 2) Even from the perspective of animal welfare studies there are disparities between treatments of sentient animals with higher and lower cognitive abilities. Therefore, it might be important to determine which level of cognitive abilities is exactly in play cases where 'cognitive bias' is found. 3) Pinpointing underling cognitive abilities in different species might clarify minimal requirements of cognitive and emotion-like system to produce such a phenomenon. This is important because evidence of "cognitive bias" across the animal kingdom fueled a heated debate of attributing emotional states and consciousness to species that are usually not considered being "sentient" (Mendl and Paul 2016) – a debate that has ramifications for animal welfare and animal rights.

2 Epistemic limits of emotional indicators

There are two major types of problems with emotional indicators like behavioral and physiological changes. First, they are not unique to specific emotional states. In other words, two or more different emotional states could be accompanied by the same/similar physiological and behavioral changes. This means that the indicators are not always indicators of *uniquely one* emotional state. Let us call this type of problems *the specificity problem* of emotional indicators. This is especially problematic in animal welfare science, for it is crucial in this field to assess specific emotional states, or at least to be able to determine if an emotional state is negative or positive for the animal. For example, detection of elevated level of circulating glucocorticoids as compared to the baseline could indicate that animal is in a

negative state (e.g., fear), but the same effect would be observed if the animal is aroused positively and thus in a positive state (e.g., reward anticipation). Without appropriate context, the elevated glucocorticoid level thus turns out to be an indicator for emotional arousal in general rather than indicating a negative state. (Ralph and Tilbrook (2016). Or, to take another example, play behavior is generally considered a good indicator of positive emotional states, but in some cases, increased playing activity was connected with a negative emotional state of the tested animal (Richter et al. 2016); Ahloy-Dallaire et al. 2018).

The second type of problem concerns the reliability of the emotional indicators *as indicators*. The observed physiological and behavioral changes are not exclusively caused by emotional states. A specific change in an animal's behavior or physiological state could be caused by, for example, by an adaptive coping mechanism that does not involve any emotional states rather than by an emotional state. Moreover, certain stereotypic behaviors are unreliable indicators of emotional states because they could be "do-it-yourself-enriching" of the environment or by calming themselves; It has been shown that stereotypic behavior could be a direct way to cope with a stressor (e.g. poor housing conditions) and thus blocking stress directly rather than indirectly *via* first eliciting another emotional state that then lowers stress (Mason and Latham 2004). Let us call this type of problems *the reliability problem* of emotional indicators.

One way to overcome these epistemic difficulties is to look for new ways of assessing animal emotional states that are (1) more emotion-specific, (2) more reliable or (3) give rise to more reliable and/or emotion-specific indicators *in combination with* already existing indicators. Before looking at how 'cognitive bias' experiments, as a relatively new and popular emotional indicator, is handling these problems, let us see how exactly the experiments that show this connection are set up.

The cognitive bias experiments are generally designed to show that 'decision making' and judgment of non-human animals are influenced by their emotional states. The rationale of the "cognitive judgment bias" test can be described as follows: Animals are first trained to respond differently to two distinct cues: one response to the "positive" cue to obtain a reward, and a different response to the "negative" cue to avert punishment (the *training phase*). When the animals have learned to respond correctly, they proceed to the test in which they are presented with "ambiguous" cues (the *testing phase*). These cues are qualitatively between the ones which were associated with negative and positive effects – hence "ambiguous". The behavioral response to ambiguous cues is considered to indicate whether animal 'anticipates' positive outcomes (responding as expecting reward) or negative outcomes (responding as

avoiding punishment). These responses are shown to be sensitive to a change in emotional state and they can ultimately be used as indicators of emotional states. For interpreting the observational data it is required to know in which emotional state they are when responding to the ambiguous cues. Therefore the animal is manipulates in a (emotionally) *priming phase* to be in a certain emotional state before being tested.

A typical setting is dividing animals into two groups. One group will be manipulated by a treatment considered to induce a negative emotional state. The other serves as the control group and would not be manipulated. Priming could be an "unpredictable" housing, lasting throughout the training and testing phase of the experiment, or an enforced electrical shock applied just before the testing phase.

In the test, animals that are primed "negatively" and thus are in a negative emotional state respond more often in the negative way, i.e., by the behavior they have learned to avoid punishment, than animals from the control group. Similar experiments are done using "positive" treatment (e.g., Matheson et al.; 2008, Richter et al. 2012; for a review cf. Lagisz et al. 2020).

Let us now come back to the epistemic problems with emotional indicators. It seems quite obvious that cognitive bias inherits the reliability problem: The described experiments may show that there is a correlation between some emotional states and judgment/cognitive bias. It would however be fallacious to assume that every case of cognitive bias is caused by some emotional state. And if it is possible and plausible that cases of cognitive bias could occur without any involvement of emotional states, then cognitive bias has the same reliability problem as other indicators. This, of course, does not mean that the overall reliability could not be increased if we took additional indicators into account. The point is rather that if we look at each emotional indicator (including cognitive bias) separately and try to assess the emotional states by it, then the reliability problem remains undissolved and *is* in fact problematic for empirical research.

At first glance, it seems that the specificity problem, too, accompanies cognitive bias as an emotional indicator, for it is hard to image that one could be able to specify the exact kind of the emotional state of an animal (fear, depression, anxious, worry, frustration, etc.) just from the judgment bias, be it negative or positive. However, the experiments seem to suggest that there are correlations between negative bias and negative emotional states in general and between positive bias and positive emotional states in general, so that the exact kind of state might not be relevant. This is certainly relevant and might in many cases even be sufficient from the perspective of animal welfare science, because, as mentioned before, the particular

interest is assessing whether or not animals are in negative (or positive) emotional states. So while, for example, an increase in heart rate could be indicating either a state of fear or one of excitement and thus does not allow to infer a negative or a positive emotional state, a state of fear would usually correlate with a negative cognitive bias and a state of excitement with a positive one. In this respect, cognitive bias promises to turn out being more specific than other indicators. It needs to be established, however, whether specificity extends to *all* positive and to *all* negative emotional states, respectively.

To sum up, in light of inherent epistemic problems of emotional indicators there are (at least) two reasons to consider cognitive bias as an alternative: (1) where emotional indicators have different degrees of reliability, having more indicators *in addition to* the already existing ones can increase the overall reliability of these indicators when they are pointing to the same emotion. (2) a cognitive bias test – if successful – seem to assess whether the indicated emotion is negative or positive, which would be of eminent value for animal welfare science.

Having discussed some inherent problems with emotional indicators and established the epistemic motivation of cognitive bias tests, let us now consider the underlying cognitive mechanism.

3 Underlying cognitive abilities/processes

3.1 Possible candidates

Scientists experimenting on cognitive bias often do not ask the question about the (exact) kind of cognitive abilities that brings the bias about and are very cautious in classifying the responses merely being "as if" the animal expected a certain outcome (Lagisz et al. 2020).. They usually treat the involved cognitive mechanism as a black box and track it though its behavioral outputs.[2] As clarified before, we think that this question is worth answering from both, the perspectives of cognitive science and of animal welfare science.

Our approach to answer the above question is to make a list of cognitive abilities and/or processes that are discussed in philosophy of cognition and that we at the same time consider being evolutionary plausible candidates that might produce the biased output in a systematic or regular way. This will outline some of the possible and plausible underlying abilities/processes within the assumed black box. It requires conceptualizing the states of the

[2] Mendl et al. (2009) sketch a picture of what they hypothesize as underlying mechanisms of 'cognitive bias' which we will in part discuss in this section. However, they admit that this might not concern animal welfare studies in practice: "From a practical animal welfare perspective it is perhaps not necessary to understand the processes underlying judgement biases" (ibid. 172).

mechanism as representing the external cues, i.e., the states are taken to indicate, to stand for, or to substitute the cue in internal processes (e.g., Sterelny 1990). For the sake of simplicitly, a representation or a cue can be taken to be the inner picture of the cue – without this metaphor meaning that representations are necessarily pictorial, or that they needed to be conscious.

1. Misrepresentation. One of the most plausible scenarios that could hold is that the ambiguous cue is represented – wrongly – as one of the cues the animal was trained upon, i.e., that it is misrepresented (Dretske 1986; Godfrey-Smith 1989). Assume the cues trained upon were squares and circles, and the ambiguous cue being an octagon. If the "inner picture" is an octagon (however one could possibly find this out), the ambiguous cue would be represented correctly (Neander 2006). If the ambiguous cue is represented either as a circle or as a square, it is misrepresented. Or consider the following standard example: of a misrepresentation (Agar 1993): a frog f_1 *mis-represents* a certain black particle, let's say a small black piece of paper, in the air as a nutritious flying prey (or something of the kind)[3], and the prey-capture mechanism of f_1 triggers a tongue-dart in the appropriate direction and captures the piece of paper. This could happen for various reasons: the black piece of paper is just too much like a fly or the frog is just too hungry etc. The point is that the piece of paper is not represented as a piece of paper (which would be impossible as long as we assume that this category does not exist at all for the frog), and that it is also not the case that it is not represented at all. It is represented as something else with which the frog is familiar with, in this case as a fly. It is likely that something similar is happening when an animal observes an ambiguous cue: that the cue is misrepresented as a familiar one.

Although Mendl et al. admit that something like this might be the case by near-positive and near-negative cues – because these cues might be too much like the positive and negative training cues (Mendl et al. 2009, 172) – they assume that it is likely that in case of perceiving the middle ambiguous cue something cognitively more advanced like 'decision-making' is happening. They base their assumption on the fact that there is evidence (at least in some studies) that the animals – or their perceptual apparatuses – are able to discriminate clearly between the middle ambiguous cue and the other training cues (Mendl et al. 2009, 173). This seems to imply that the ambiguous cue is represented in another way than any of the trained cues and consequently that it is not misrepresented as one of these. However, this would be too quick a conclusion. Although we grant that something like this might be happening in

[3] In this paper we do not want to discuss what it is exactly that the frog represents, i. e. how is the content of the (mis-)representation fixed. For more discussion on this matter see Schulte (2012).

animals with higher cognitive abilities, which we will consider next, we want to emphasize misrepresentation being one of the most likely scenarios. To be clear, our estimation of likelihood here does not foot on empirical data but rather on the principle of Ockham's razor to be as scarce as possible with assuming entities, in this case: with presupposing involved cognitive instances or abilities. Consider the following: just because one is able to distinguish between cats and dogs under ideal or under standard conditions, it does not mean that one is not likely to confuse them under certain circumstances or in certain contexts, e.g., to mistake in dim light a small dog for a cat. Similarly, just because animals investigated in a cognitive bias test are able to discriminate between the middle ambiguous cue and the cues in the training phase under certain conditions, they need not be able to do so under the conditions of the discrimination experiment. They might still misrepresent the middle cue as one of the training cues. In fact, if it is likely that the animals are misrepresenting the ambiguous near-positive and near-negative cues, we do not need to – and should not – bring some higher cognitive abilities, like 'decision-making', into play to explain their response to the ambiguous middle cue, even if from the perspective of humans that was what we would usually do when we interpret an ambiguous cue. We judge misrepresentation likely because there could be so many reasons for it that make it plausible: the reward is just too delicious, or at least delicious enough to mistake anything *resembling* the positive cue as *being* the positive cue; or the punishment is too severe or severe enough so that anything resembling the negative cue gets mistaken as being the negative cue; or the emotional inducing phase made the test animals too cautious, too afraid, too anxious, too bored etc. Therefore, applying the Ockham-razor-principle, we assume that we do not need to use more advanced cognitive abilities to explain the responses to the ambiguous cues (middle or near ones) as long as there is no further evidence that suggests the involvement of such abilities. However, because it is possible that more advanced cognitive processes would produce the similar output under the similar input conditions (as it is possible in the case of human animals), we will still consider this option and try to identify the minimal requirements of such a cognitive system according to an evolutionary perspective.

Before we introduce this option of the involvement of a decision mechanism, let us illustrate another possible candidate, which would be empirically discernable from the other two options:

2. Constitutive lack of discrimination. We think that it is plausible (but rather easy to rule out in the described experiments, depending on the kinds of cues being used) that the cognitive system of some animals does not discriminate between the cue that, *from our*

perspective, should be ambiguous for them, and one of those that are associated with positive or negative effects. This inability is rather a 'constitutive' lack of discrimination, one, that is not mediated or altered by emotions and other conditions, for it is a matter of physiology and unmodifiable by priming. Imagine, for example, somebody who suffers from a particular kind of color blindness and cannot discriminate between, say, blue and purple (but can tell red!). This person now receives a purple cue, meant as middle cue between blue and red and, and sees it blue. The test person's perceptual apparatus does simply not discriminate between what we would classify as a middle cue and as one of the others. Now imagine that this is the 'normal' case for the whole species that is being experimented on. This possibility could be eliminated in the cognitive-bias-experiments by conducting a separate experiment that shows whether or not the animals are able to distinguish between the different cues that are being examined. This option does not interfere with the interpretation of most of the experiments, since the lack of discrimination between training cue and supposedly ambiguous cue can be easily detected. We are mentioning it merely for reasons of completeness, and because it helps to better understand the other candidates.

3. Conflicting content(s). The third possibility which could be available in an advanced cognitive system is the representation of the ambiguous cues *as ambiguous*, for example as something undetermined between two or more *specific* states or objects. To have an analogy from the perspective of a (human) viewer, it is not like: "I am seeing something but I don't have any idea what it is", but like "I am seeing something that is either *x* or *y*, but I cannot exactly tell which of those two.".

Each of these three cases might be considered an ambiguous representation. But the latter is analog to the cases that we are considering. It is also important to note that the conflicting content(s) could be different contents of different representations of the same state of affairs, or a 'conflicting' content of one representation of that state of affairs. Without going too deep into the theories of content, with a *conflicting* content of one representation we are referring to a content that has two or more aspects with different psychological roles (hence 'conflicting'), e.g. a state of affairs is represented as being a dog or/and a cat or even as a dog or/and a non-dog, where there are different behaviors associated with these different aspects, for example fleeing in case of the representation of a dog and attacking in case of a cat or a non-dog. How exactly these aspects are represented and how the connections between them look like are not relevant here. Relevant is here only that the cognitive system links these different aspects to

different behavioral outputs[4] and that the cognitive system has means to deal with this conflict.

While it might sound natural that humans have such representations, the issue is much more complex than it appears at first glance. In general, the state of affairs in question needs to be represented *as 'conflicting'* (either through the conflicting representations or the conflicting aspects of a representation of the state of affairs), which furthermore means that there are mechanisms, over and above 'regular' representational mechanisms, that evaluate these representations and compute, or 'decide' about,[5] the generation of an output signal that enters the behavior-producing mechanisms. This feat of the cognitive system is over and above the ability to represent (and misrepresent) something in a specific way, because simple representational systems do not usually evaluate representations or aspects of a representation *against each other*.

We want to emphasize that we are not suggesting that there is no evaluation of representations or some kind of computing happening in cases of mere misrepresentations. However, if the animal has an 'ambiguous' representation, then it probably has competing representations and some kind of 'resolving'-mechanisms that deals with the 'ambiguity'. And this seems to be a different, more advanced cognitive ability or process than merely representing a cue. Bear in mind that from the set-ups of the experiments there is not yet much known that allows us to assess which kind of these cognitive abilities or processes (misrepresentation *versus* conflicting contents) is in play. Our analysis suggests a way of gaining better knowledge about the representational systems, i.e., a way to open the black box at least a little bit: does the animal react always the same way to an ambiguous cue, or does it learn to discriminate? One might expect that conflicting content is interpreted cautiously or with hesitation on the first confrontation, but more decided in later ones, while a plain misrepresentation would not give rise to any hesitation.

Such experiments, however, would merely hint at certain mechanisms, since it is hard to see how comparison with a control organism could be done in a valid way. We will therefore discuss more complex experiments that could yield more definite results on the representation mechanism involved.

[4] "Behavioral output" is to be understood in a broad sense and does not need to be a behavior of the organism. It includes, for instance, activities of some subsystems that are triggered by the representation(s).
[5] Here "computing/'decision-making'" does not need to be a conscious process.

3.2 Ways of differentiating: A new proposal

As we stated earlier, the possibility that the tested animals might lack the ability to discriminate between the ambiguous cues and the cues in the training phase can be eliminated through separate experiments that tests their perceptual abilities. However, things are more complicated if we were to establish whether a behavioral output of the cognitive-bias-test is the result of a misrepresentation or of an 'ambiguous' representation, in particular in cases where the observation of the behavior in subsequent experiments, as proposed above, does not allow deciding among the possible cases.

As a promising way to eliminate the possibility of misrepresentation in the cognitive-bias-test we propose a setting that does not involve an 'ambiguous' cue at all. We suggest to use two pairs of cues instead, that might even address different sensory modes. 'Ambiguity' (or conflict) could then be realized by combining the positive cue of one of the pairs with the negative cue of the other. In the training phase, animals needed to learn, over and above the content of the training phase in the standard setting, associating the two additional sensorially different cues with negative and positive outcomes, respectively. In the testing phase, the properly conditioned animals will be exposed not to one ambiguous cue, but rather to two sensorially different cues[6] simultaneously. One cue is associated with a negative and the other with a positive outcome. Each of the cues is unambiguous.[7]

3.3 Possible outcomes

In the following, we discuss the possible outcomes of such an experiment and show which conclusions could be drawn with respect to how the underlying cognitive system represents the cues:

It is important to test both options of ambiguous combinations of cues, a positive cue 1 with a negative cue 2 and a negative cue 1 with a positive cue 2. This rules out that one of the cues might generally override the other. Only consistent answers to both ambiguous combinations allows ascribing emotional bias.

As the first setting let us assume that the experiments are done without the priming phase. The individuals thus are trained to the cues and then exposed to an ambiguous combinations of cues, without any prior exposition to emotion-eliciting conditions.

[6] Optimally, both senses should have similar perceptual values for the animals to avoid the possibility that the behavioral outcome is the result of the animals being overly sensitive to one cue rather the other.

[7] Of course there should be several control groups with negative-negative, positive-positive and negative-positive with different timely distance between the sensorially different cues.

(ai) Each individual might show a consistently biased answer, positively in some individuals and negatively in others. This would allow ascription of a stable dispositions to the individuals that count as long-lasting emotional states, or individual trait-emotions, in the literature usually called ‘optimism’ and ‘pessimism’ (Krakenberg et al. 2020).

(aii) All individuals might show the same bias, either positive or negative. One could interpret this as constitutive optimism or pessimism being trait-emotions of the species under investigation, where either a positive or a negative cue overrides an opposing cue.[8] Such ‘bold’ or ‘cautious’ species trait emotions might be selectively advantageous under certain living conditions so that their existence could be expected. They might blend with individual variable trait-emotions.

(aiii) The answer might be found to be arbitrary in all individuals, i.e., the ambiguous combination of cues leads to positive and negative answers in statistically indiscernible proportions in each individual. The conflicting contents, which in isolation lead to a positive and negative answer, respectively, level out. No emotional bias can be found. This means that the pre-treatment did not evoke any emotion that lasts until application of the ambiguous combination of cues, or that the bias is too small to be detectable by means of the performed experiments. It does not rule out that a modified or refined experiment might indicate emotional bias, e.g., one using different cues or, where this option holds, cues of different intensity.

As a second setting let us consider experiments that involve the priming phase as described above in Section 2. The animals that are trained to the cues are primed for positive or negative emotions before exposing them to ambiguous combinations of cues (reciprocal combinations required, as in a). The prime target of such experiments is the influence of variable emotions or emotional episodes, so-called state emotions, on behavioral responses to ambiguous combinations of cues. We are therefore discussing only possible outcomes of experiments addressing this bias in general, where individual differences are manifest only in statistical parameters. Experiments with ambiguous combinations of cues can address individual differences of state-emotions as well, which requires blending the a- and b-evaluations.

(bi) The animals might show a positive emotional bias with respect to the ambiguous combination of cues after certain kinds of priming, and a negative bias after other kinds. This would allow ascribing treatments resulting in positive bias to evoke positive state-emotions, and treatments resulting in negative bias to evoke negative state-emotions.

[8] Please keep in mind that we take emotions being dispositions that need not be conscious.

(bii) It would be possible that all pre-treatments evoke positive bias, or that they all evoke negative bias. If it could really be established that such a result is not artifactual, the experiment might simply have missed to evoke in phase 2 negative and positive state-emotions, respectively – perhaps because of a lack of knowledge of the living conditions and demands of the investigated species. It would also be possible that *any* change of the emotional state of an individual of the species under investigation is positive or negative, respectively. However, we would not expect any species showing such constitutive negativity or positivity of any emotional episode or state-emotion.[9]

(biii) Pre-treatment might also not lead to any change in the interpretation of conflicting representational content, so that no emotional bias can be found. This might mean, as in aiii), that the pre-treatment did not evoke any emotion that lasts until application of the ambiguous combination of cues, or that the bias is too small to be detectable with the concrete experimental setting. However, as long as data from different individual are averaged, it might also be the case that individuals react on the treatment with different emotions, some with positive ones and other with negative ones.

Conclusion

Cognitive bias tests of judgment allow assessing emotional states of non-human animals. Central to these tests is confronting animals with ambiguous cues that are intermediates between cues they have learned to link to positive and negative consequences, respectively, and to act accordingly. The mechanism of decision-making is usually taken to be a black box. We discussed how this black boxed could be opened at least a little bit even by experiments of the considered type. Drawing on the philosophical perspective of understanding decision making as a capacity of certain representational systems, we determined three different ways how ambiguous stimuli could in principle be represented. We propose that a test regime in which the ambiguous stimulus is replaced by an ambiguous pair of unambiguous stimuli. The discussion of the various possible outcomes of such experiment suggests that in many cases the way in which the ambiguous combination is represented could be inferred. This would not only be an interesting result in itself, but also help better understanding the mechanism of biased judgment in non-human animals and to develop further experimental tests of cognitive bias.

[9] Imagination of such situations might be advanced by pictures that involve conscious experience of emotions in humans: Even joy, though felt positively, can evoke negative expectations in a life scheme that aims at complete peace of mind, at Epicurean *ataraxia*. And even suffering, though felt negatively, can evoke positive expectations in religious framings that damn joy, and perhaps even promise rewards in an afterworld.

Funding

This research was funded by the German Research Foundation (DFG) as part of the SFB TRR 212 (NC³) – Project number 316099922.

References

Agar, N., [1993]: 'What do Frogs Really Believe?', *Australasian Journal of Philosophy*, **71**, pp. 1-12.

Ahloy-Dallaire, J. and Espinosa, J. and Mason, G. [2018]: 'Play and optimal welfare: Does play indicate the presence of positive affective states?', *Behavioural Processes*, **156**, pp. 3-15.

Dretske, F. [1986]: 'Misrepresentation', in: Radu Bogdan (*ed*.): *Belief: Form, Content and Function*, New York: Oxford University Press 1986, pp. 17-36.

Godfrey-Smith, P. [1989]: 'Misinformation', *Canadian Journal of Philosophy*, **19**, pp. 533-550.

Harding, E.J., Paul, E.S. and Mendl, M. [2004]: 'Animal behavior—Cognitive bias and affective state', *Nature*, **427**, p. 312.

Krakenberg, V., Siestrup, S., Palme, R., Kaiser, S., Sachser, N. and Richter, S.H. [2020]: 'Effects of different social experiences on emotional state in mice' *Scientific Reports* 10:15255, https://doi.org/10.1038/s41598-020-71994-9

Kremer, L., Klein Holkenborg, S., Reimert, I., Bolhuis, J. and Webb, L. (2020). The nuts and bolts of animal emotion. *Neuroscience & Biobehavioral Reviews*, *113*, 273-286. https://doi.org/10.1016/j.neubiorev.2020.01.028

Lagisz, M., Zidar, J., Nakagawa, S., Neville, V., Sorato, E., Paul, E.S., Bateson, M., Mendl, M. and Løvlie, H. [2020] Optimism, pessimism and judgement bias in animals: A systematic review and meta-analysis, *Neurosci Biobehav Rev* 118:3-17, https://doi.org/10.1016/j.neubiorev.2020.07.012

Mason, G.J. and Latham, N.R. [2004]: 'Can't stop, won't stop: is stereotypy a reliable animal welfare indicator?' *Animal Welfare*, 13, pp. S57-S69.

Matheson, S., Asher, L., and Bateson, M. (2008). Larger, enriched cages are associated with 'optimistic' response biases in captive European starlings (Sturnus vulgaris). *Applied Animal Behaviour Science*, *109*(2-4), 374-383. https://doi.org/10.1016/j.applanim.2007.03.007

Mendl, M., Burman, O. H. P., Parker, R. M.A. and Paul, E. S. [2009]: 'Cognitive bias as an indicator of animal emotion and welfare: Emerging evidence and underlying mechanisms', *Applied Animal Behaviour Science*, **118**, pp. 161-181.

Neander, K. [2006]: 'Content for Cognitive Science', in G. McDonald and D. Papineau (eds.), *Teleosemantics*, Oxford: Oxford University Press, pp. 167–194.

Neville, V., Nakagawa, S., Zidar, J., Paul, E., Lagisz, M., and Bateson, M. et al. (2020). Pharmacological manipulations of judgement bias: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, *108*, 269-286. https://doi.org/10.1016/j.neubiorev.2019.11.008

Paul, E. S., Harding, E. J. and Mendl, M. [2005]: 'Measuring emotional processes in animals: the utility of a cognitive approach', *Neuroscience and Biobehavioral Reviews*, **29**, pp. 469–491.

Ralph, C. R. and Tilbrook, A. J. [2016]: 'The usefulness of measuring glucocorticoids for assessing animal welfare', *Journal of Animal Science*, **94**, pp. 457–470.

Richter, S. H., Kästner, N., Kriwet, M., Kaiser, S. and Sachser, N. [2016]: 'Play matters: the surprising relationship between juvenile playfulness and anxiety in later life', *Animal Behaviour*, **114**, pp. 261-271.

Richter, S., Schick, A., Hoyer, C., Lankisch, K., Gass, P. and Vollmayr, B. (2012). A glass full of optimism: Enrichment effects on cognitive bias in a rat model of depression. *Cognitive, Affective, & Behavioral Neuroscience*, *12*(3), 527-542. https://doi.org/10.3758/s13415-012-0101-2

Sterelny, K. [1990]: *The Representational Theory of Mind: An Introduction*, Cambridge, MA: Blackwell.

Schulte, P. [2012]: 'How Frogs See the World: Putting Millikan's Teleosemantics to the Test', *Philosophia*, **40**, pp. 483-496.